

Machine Intelligence in Healthcare: Perspectives on Trustworthiness, Explainability, Usability and Transparency – Executive Summary

July 12, 2019, Neuroscience Center, National Institutes of Health, Bethesda, Maryland

BACKGROUND

Machine Intelligence (MI) is rapidly becoming an important data science-based analytic tool across biomedical discovery, clinical research, medical diagnostics and devices, and precision medicine. This meeting defined MI as the ability of a trained computer system to provide rational, unbiased guidance to humans in such a way that achieves optimal outcomes in a range of environments and circumstances. MI systems have the potential to allow physicians and patients to make more informed care decisions and to achieve better medical outcomes through improved analysis of medical images, use of medication adherence apps and clinical decision support tools, among others.

Current questions in the field are: how do we trust MI system outputs and decisions without any attached context or fit statistic? How do we ensure that MI outputs are safe and beneficial for human health? Do we understand how the adaptive nature of MI systems can change output? These questions are especially relevant to clinical care decision making – are the risks of using such systems understood and how can the technology be deployed for maximal benefit in healthcare settings?

GOAL

To provide experts an opportunity to share their perspectives on current issues associated with incorporation of MI systems into healthcare settings. Meeting outputs will be used to develop a whitepaper on translating MI for clinical applications and the associated process improvements needed when implementing MI systems in healthcare environments.

In the context of this meeting, the following additional definitions were used:

Trustworthiness – the ability to accept the validity and reliability of a result, given a change in input or algorithmic parameters, without necessarily knowing how the result was derived. Healthcare professionals need to be able to determine when a result is wrong (or the probability that a result is wrong) and ensure that the result is interpreted correctly without needing to know what happened “under the hood”.

Explainability – the ability to understand and evaluate the internal mechanics of a machine or deep learning system in human terms. As these MI systems are being built, additional steps will need to be added in the development process to account for: data quality, metrics for the system’s functioning and impact, standards for applications in the healthcare environment, and future updates to the data/system.

Usability – the extent to which an MI system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in multiple healthcare environments. How useful will these systems be to both doctors and patients in multiple settings, particularly when compared to systems and standards of care that are already in place?

Transparency and Fairness – the right to know and understand the aspects of a dataset/input that could influence outputs (clinical decision support) from algorithms. Such factors should be available to the people who use, regulate, and are affected by the systems that employ those algorithms. For MI systems in healthcare, we will need to identify ways of evaluating and preventing bias, encouraging data transparency, and ensuring open access to MI system development, each of which have unique challenges associated with them in the context of healthcare.

EXECUTIVE SUMMARY

Session 1 – Trustworthiness

Trust that the results/actions of an MI system are reliable and valid requires that we trust all steps (from end-to-end) of its workflow. However, the meaning of trust can change depending on focus, i.e., a machine learning expert may define trust in a system according to different parameters than a user. On a high-level, trust is proof over time that a system does what it claims to do – this can only be gained over time and through continual monitoring (via use cases and output validation in comparison to standard of care). Trust is essential for *uptake* of an MI system – for all users, but importantly for the patient and provider. To improve uptake, clinicians need: sufficient warning prior to system updates, transparency and understanding of what the output means, publicly available benchmarks to rate outputs, and expert review of outputs to ensure consistency of the system.

- We need to be able to trust the data –
 - In some ways, it could be argued that there are no “bad data,” just a question of whether the data are a good fit for the system’s purpose. Since many MI systems are built on secondary data sources, “scoring” the data and its fit/quality/relevance for the MI system is integral.
 - In order to mitigate the issue of “garbage in, garbage out,” developers could limit the “reference standard” for MI systems to data sources that meet specific criteria or are above a cut-off.
 - To track progress of interventions across the board, there is a need for a good approximation of current evidence and further tracking over time.
 - Particularly in the healthcare space, there is a need to extract structured data from unstructured information held within health care systems in an automated way.
- We need to be able to trust the system –
 - Reproducibility, replicability, and robustness of the technical and conceptual aspects of the MI system should be tested.
 - System frameworks can be utilized, but transferability of an MI system across disparate localities is likely not feasible on a broad scale – for example, bridges are widely used, but need to be modified for specific locations.
 - MI systems and algorithms do not provide causation, only associations – randomized trials are needed for systems to determine causation.
- We need to be able to trust the system output – An additional system (can be human) should evaluate trustworthiness of the system’s output, assessing whether the test and training data are anomalous to that of the output data.
- We need to be able to trust the workflow – Continual surveillance of the MI system is essential even after the system is in use – just like in other industries, there is a need to monitor long-term outcomes and establish structured feedback loops (i.e., post-market surveillance).

Session 2 – Explainability

Explainability boils down to translation between machine and human logic in a way that a human can understand. MI examinability is often geared towards answering a “why” question and can be broken down into two contrasting types of explanations – interpretability (describing the internals of a system in a way that is understandable to humans) and completeness (describing the operation of a system in an accurate way). Different levels of explainability may be needed depending on the user and desired depth of explanation. Importantly, however, context matters; and correlation is not causality.

- Explaining the processing of data – Examples include proxy models (an interpretable system that behaves analogously), occlusion visualization, saliency maps, and class activation maps.
- Explainability methodologies – The creation of explanation-producing systems through network dissection, disentangled representations and explicitly training networks to generate explanations.

Explainability is again crucial for *uptake* of an MI system, so it is necessary to express and share how the system operates with key stakeholders. To do so, some methodologies include: exposing the training data and labels used for system development; applying techniques to increase visibility of inputs that indicate how the system determined the output; forming a dialog with stakeholders on how the system works; illustrating the output as a score assisted by visualization (an explanation interface); and educating stakeholders on the tradeoffs for false positives/negatives for prediction and inference in the system. The panelists also emphasized the need for actionable insights – prediction itself is not an intervention and a prediction does not imply causality.

An important challenge within MI is the creation of crisp definitions so that practitioners and users in different fields can communicate more meaningfully. Additionally, it is important to note that MI systems are tools, not “intelligent” systems on their own – therefore, it is crucial that the clinician be empowered to navigate the landscape – when the clinician and algorithm disagree, the clinician should be equipped to determine the best path forward. Overall, explainability is particularly important at our current point in time in order to encourage acceptance of these systems.

Session 3 – Usability

There are many types of patient-generated data sources with direct applicability to healthcare – e.g., medication history, demographics, patient-reported symptom assessment, social media, wearables – but it is difficult to integrate and interpret them at the point-of-care. Despite major investments in electronic health records (EHRs), the point-of-care setting remains a “walled garden” due to isolated, non-interoperable, and tailored implementation across health care provision sites. Current work around this issue includes a programming interface layer that sits on top of the EHR system, which runs across individual instances of the EHR to connect health systems data. However, even with such a system in place, patient-generated data will most likely still remain non-standardized and siloed. It is possible that, in the future, MI systems may be able to further assist with this issue by pulling reference population data from EHRs into standardized analytic platforms to enable population health efforts. Key lessons within the context of usability:

- System feedback is a key factor.
- There needs to be a plan for unforeseen consequences and circumstances.
- Interfaces need to be user centered.
- There is a need to understand the current workflow and its constraints.
- Interventions should be compared on the basis of health-related outcome measures, i.e., comparative effectiveness of MI systems against standard of care should be considered.

Session 4 – Transparency and Fairness

Issues related to ethics, transparency, and fairness in healthcare are not new. However, new issues arise and are especially important within MI applications in healthcare. There is the challenge of bias or unintended weighting, its amplification, and its effects on system outputs. Even with training, bias (whether it be introduced by humans, algorithms, or data) remains a part of the clinical landscape. How can we mitigate bias perpetuation when designing these systems? Some suggested solutions were:

- Humans – tracking and education regarding biased language.
- Algorithms – development of approaches to identify bias in results.
- Data – utilization of representative data sets.

The ultimate goal should be processes with transparent frameworks rather than transparent systems, as it is impossible to completely audit the data, and “big picture” tools should be developed to better understand potential biases in a data/model-agnostic manner.

MI systems require large amounts of data, incentivizing the collection of it on a large scale – requiring the identification and collection of good data, while addressing privacy issues. Additionally, MI systems function at a high-intelligence level, but will not apply ethical principles to how the data is used (unless explicitly trained to do so). This emphasizes the need for transparency in the systems, so that users can interpret the internal decisions that have produced the output. However, some argue that interpretability is less important than the ability of the system to do a “good job.” MI systems cannot determine causal relationships and accuracy/reliability may be more important than explainability itself.

Within the area of transparency and fairness, thinking about MI applications in terms of equality of healthcare, treatment, and outcomes is crucial:

- Equality in healthcare/access to healthcare – Better healthcare can lead to better predictions about health; those with poorer or no healthcare will have a harder time receiving analogous predictions.
- Equality of treatment – If two patients go in to see the doctor, both should receive the same quality and level of treatment; this is currently not the case.
- Equality of outcome – Two patients get the same drug, but they respond differently. Equal outcomes may require a different treatment(s).

We must ask – where should ethical approaches be emphasized and what outcomes should be optimized to most effectively address bias? Incorporation of not just health data, but also social determinants of health and health outcomes, into the overarching pool will aid in chipping away at inequality issues.

Finally, the panel touched on the issue of risk and the application of MI systems to maximally benefit society. Lots of patient and population data already exist – if those data can be appropriately pooled, clinicians might be better able to intervene in health. However, while one could make the argument that there is an obligation to put the group’s needs over the individual, this could compromise the safety and privacy of the individual.